Slide 1 of 37

PIAAC Sample Design, Weights, Variance, and Missing Data

Slide 2 of 37

Module Objectives

- Describe the PIAAC sample design
- Describe the weights that must be applied to ensure data are representative of the target population
- Describe the procedures for calculating appropriate standard errors
- Explain how scaling is used in the large-scale international assessments and how plausible values are used when analyzing assessment data
- Describe how to account for missing data to ensure accurate analysis

Overview of Module Topics

- Target Population
- Sample Design
    - o Frame, sampling units, sample size, eligibility criteria, quality standards
- Assessment Design
    - o Quality standards, overview, eligibility criteria
- Sampling Weights
    - o Non-response and other adjustments, the final weight, and corresponding replicate weights
- Proficiency Estimation Plausible Values (PVs)
    - o Estimation of the statistic and variance using PVs
- Missing Data

Slide 4 of 37

PIAAC Target Population

- All non-institutionalized adults, ages 16-65, who reside in the country at the time of data collection, irrespective of nationality, citizenship, or language status
    - o Residence: Living in a house, an apartment, or a condo as well as workers' quarters, halfway homes, and dormitories, fraternities, or sororities
    - o Non-institutionalized adults: People not in prisons, hospitals, nursing homes, or military barracks/bases
    - o Countries allowed to oversample and include additional subpopulations of interest (not in original sample)

Slide 5 of 37

PIAAC Sample Design and Weighting

- The PIAAC standard sample design was a self-weighting design of persons (or of households, for countries without person registries)
    - A self-weighting design is achieved when each sample person (or household, for countries without person registries) has an equal probability of selection
- PIAAC standards require probability-based samples
- Geographically small countries used fewer sampling stages and less clustering compared to larger countries

Slide 6 of 37

Sampling Frame

- Several types of frames were used to select the household samples
    - Lists of residents (registry)
    - Master samples used for national surveys
    - Stratified area sample designs
- Participating countries used
    - One-stage stratified or non-stratified sampling
    - Two-, three- or four-stage stratified probability proportionate to size sampling designs

Slide 7 of 37

Sampling Units

- A number of countries implemented a one-stage sample design, in which there is only one sample unit: persons. This included Austria, Belgium, Denmark, Estonia, Finland, the Netherlands, Norway, and Sweden

Country	Sampling Stages			
	Stage 1	Stage 2	Stage 3	Stage 4
Austria	Persons	--	--	--
Flanders (Belgium)	Persons	--	--	--
Denmark	Persons	--	--	--
Estonia	Persons	--	--	--
Finland	Persons	--	--	--
Netherlands	Persons	--	--	--
Norway	Persons	--	--	--
Sweden	Persons	--	--	--

Slide 8 of 37

Sampling Units (Continued)

- The sampling units for countries with two-, three-, and four-stage sample designs included districts, communities, households, municipalities and dwelling units, etc.
- Sample designs are described in more detail in the Technical Report of the Survey of Adult Skills (PIAAC) (Chapter 14, Section 14.4)

U.S. Sampling Units

- The United States used a four-stage stratified area sample consisting of the following units
    1. Primary sampling units (PSUs) comprised of counties or groups of contiguous counties
    2. Secondary sampling units, or SSUs (referred to as segments) consisting of census area blocks
    3. Housing units (DUs) containing households
    4. Eligible person(s) within households
- Person-level eligibility was determined by collecting age and gender of all household members through a screener

Slide 10 of 37

Sample Size

- Minimum sample size requirements for the standard target population depended on
    o Optional components administered in the country [problem solving in technology-rich environments (PS-TRE) and reading components]
    o Whether or not the assessment was administered in more than one language
        ▪ Larger sample sizes were needed for each additional language

Sample Size (Continued)

- Most countries administered the optional PS-TRE and reading components
- Some administered the PS-TRE only, some the reading components only, and some did neither
- Five countries performed assessment in multiple languages: Canada, Estonia, Finland, the Slovak Republic, and Spain

Slide 12 of 37

U.S. Sample Parameters

- The United States administered
    - Required literacy and numeracy components
    - Optional PS-TRE and reading components
    - English was the only language for the assessment
    - In Round 1, the U.S. target sample size was 5,000 adults
- To reach this target, 9,468 households in the U.S. were sampled and 6,916 were eligible for PIAAC
    - From these households, 4,898 adults completed the BQ and 4,820 completed the assessment

Slide 13 of 37

Eligibility Criteria

- Eligible households at the sampling stage
    - Units with occupancy (i.e., that were not vacant)
    - Units that were not for seasonal use only
    - Actual dwelling units (as opposed to institutions)
    - Units with persons ages 16 to 65
- In the U.S., the occupancy rate was 86 percent and the eligibility rate was 82 percent
- Reserve samples were selected to be used if necessary

Slide 14 of 37

Quality Standards

- To ensure high quality and high comparability of the data across all participating countries, PIAAC developed a comprehensive set of quality assurance and quality control checks for all sampling activities
    - Sample design and selection results were included
    - Most countries adhered to the [standards](#)
    - Those who did not, received cautionary remarks on the sample selection process flagging the weakness of their data comparability

---

Quality Standards (Continued)

- PIAAC also developed a comprehensive set of quality assurance and quality control checks for all assessment implementation procedures
- All countries adhered to the standards, resulting in the general cross-country comparability of the assessments

Slide 16 of 37

Overview of PIAAC Assessment Design

- Two possible modes
  - Paper-based or computer-based administration of PIAAC's cognitive items (psychometric assessment items)
- Multiple matrix sampling
  - Sampled individuals received only a subset of cognitive items
- Multi-stage adaptive testing
  - The sets of cognitive items that each respondent received were based on the respondent's previous answers
  - PIAAC is the first international comparative survey to apply adaptive testing procedures

Weighting Requirements for Valid PIAAC Estimates

- PIAAC's complex survey sample design and its complex assessment design require
    - Multiple weights to obtain valid survey estimates
    - Plausible values to obtain valid assessment estimates
- "Valid" means representative of the target population

Slide 18 of 37

Sampling Weights

Sampling weights are designed to make the data representative of the target population by
- Compensating for disproportionate sampling of subgroups and non-coverage
- Making use of known data for the population
- Minimizing biases arising from differences between respondents and non-respondents
- Facilitating the estimation of variances through the use of the replication approach

Creating the PIAAC Final Weight

- A final weight is required for all sampled persons
    - With a completed background questionnaire
    - Who could not complete the background questionnaire for literacy-related reasons, but for whom age and gender were collected
- Several steps are taken to create the PIAAC final weights

Slide 20 of 37

Creating the PIAAC Final Weight (Continued)

- First, each sampled household is assigned a household base weight to compensate for differential probabilities of selection
- Then, household eligibility and nonresponse rate adjustments are made to reduce potential biases arising from differences between respondents and nonrespondents
- These first two steps are employed only for countries that used a screener, like the United States

Creating the PIAAC Final Weight (Continued)

- Next, each sampled person is assigned a person base weight to compensate for differential probabilities of selection within the household
- Then, for countries that used a registry as a sampling frame, person-level eligibility and nonresponse rate adjustments are made to compensate for sampled persons who refused to participate; were inaccessible; had a physical disability that prohibited their participation; or, did not respond due to literacy-related reasons
  - o Because of unequal household sizes, sample designs that include the selection of dwelling units have more variability in larger weights compared to directly sampling persons from registries
  - o Therefore, large weights are adjusted by trimming them to a designated cutoff value to reduce their variability, as needed

Slide 22 of 37

Creating the PIAAC Final Weight (Continued)

- Deficiencies in the sampling frame can result in a sample that fails to cover segments of the target population in proportions representative of those same segments in the population
- Therefore, the last step in calculating the final PIAAC weight, once the adjusted household and person level weights have been combined and trimmed, is to compensate for this noncoverage by calibrating person weights to independent control totals
  - o This means that the sampling weights are modified so that the totals on specified characteristics, such as age and gender, agree with corresponding totals known for the population

Final Weight Adjustments

- Variables used in the weight adjustments (e.g., calibration, non-response)
    - Age and gender was used by all countries, as required in the PIAAC Technical Standards and Guidelines
    - Region was also used in all countries
    - In addition, the majority of countries included at least one variable related to education, employment status or nationality (shown to be correlated with proficiency)

Final Weight Adjustments (Continued)

- All PIAAC variables selected for benchmarking (calibration) were required to have reliable control totals available
- Countries used Labor Force Surveys, Censuses, Registries and other sources to calibrate PIAAC population totals to accurately represent the total population
- Control population totals used in the benchmarking process were designed to have the same definition and coverage of the PIAAC target population

Replicate Weights

- Statistical analyses are not valid unless the corresponding variance estimators appropriately reflect all of the complex features of the PIAAC sample design (e.g., stratification and clustering)
- The replication approach is used for estimating variances in analyses of PIAAC data
    - This method calculates appropriate SEs based on differences between estimates from the full sample and a series of created subsamples (replicates)

Slide 26 of 37

Replicate Weights (Continued)

- Countries participating in PIAAC used one of four different replication schemes
    - Delete-one jackknife (JK1)
    - Paired jackknife (JK2)
    - Balanced repeated replication (BRR)
    - Fay's method (variant of the BRR approach)

Standard Error Calculation in U.S PIAAC Data Replication Techniques

- Creation of the U.S. Final Weight
    - Some 15 variables were used in the weight adjustment procedures, including age, gender, region, educational attainment, employment status, country of birth, a measure of linguistic isolation, and presence of children in the home
    - American Community Survey estimates were used as benchmarks to calibrate the sample to the total population
- Final Weight
    - SPFWT0
- Replicate Weights
    - Paired jackknife (JK2) method was used for replication
    - Select the 80 replicate weights that are associated with the final weight SPFWT0 are SPFWT1 through SPFWT80

Slide 28 of 37

Proficiency Estimation - Plausible Values

- PIAAC cannot provide individual-level results, because each adult answers only a small number of assessment questions
- PIAAC provides reliable estimates of proficiency only at the national level, or at the level of large subgroups (e.g., females, employed, or college educated)
- Plausible Values (PVs) allows PIAAC dataset information to be saved at the case level to estimate proficiency at the national or subgroup level
    - Each case's PVs reflect not only that individual's performance on the small number of items s/he answered, but also the performance of similar respondents on the rest of the PIAAC assessment

Obtaining Plausible Values

- Each adult's performance on the assessment questions gets summarized in the form of a probability curve or distribution, with a value for every possible score
    - The probable score distribution is sometimes bell-shaped, but sometimes not
    - The probable score distribution is based solely on the individual's own performance on PIAAC assessment questions

Obtaining Plausible Values (Continued)

- The probability distributions for all adults in a subgroup are used to estimate that group's probability distribution
- The estimates for each group's performance are then used to weight each respondent's probability distribution
- Calculations then alternate between weighting each respondent's probable score distribution by the group's performance, and combining the weighted score locations into new estimates of group performance
- The calculations continue until the group performance stops changing
    - Neither the group performance nor the weighted student performance is known in advance, but are arrived at simultaneously

Obtaining Plausible Values (Continued)

- To avoid having to perform all these calculations each time one wants to analyze PIAAC proficiency data, the results of the iterative process are saved for each respondent in the form of plausible values (PVs)
    - Each value is a plausible representation of the performance of that kind of individual
    - PVs are samples from the final weighted probable score distributions for each respondent
    - PVs reflect not only that individual's performance on the small number of items s/he answered, but also the performance of similar respondents on the rest of the PIAAC assessment

Slide 32 of 37

Plausible Values Review

- Plausible values (PV) are a statistical means to replicate a probable score distribution that summarizes
    - How well each respondent answered a small subset of the assessment items, and
    - How well other respondents from a similar background performed on the rest of the assessment item pool
- Each individual case in the PIAAC dataset has a randomly chosen set of ten plausible values (PVs)
    - All ten PVs must be used together to estimate proficiency, or else one understates the variability in the predicted outcomes
    - The randomly chosen set of PVs best represents the score distribution for a subgroup of adults
- For information on the IRT (Item Response Theory) scaling and the population model, refer to Chapter 17 of the Technical Report of the Survey of Adult Skills (PIAAC)

Proficiency Estimation

- For estimations involving proficiency scores, calculations must account for both the sampling error component and the variance due to imputation of proficiency scores
- To account for the sampling error component, use the final weight and the corresponding 80 replicate weights
- To account for the imputation variance, use all ten plausible values

Slide 34 of 37

Missing Data

- Background Questionnaire
    - All missing data are marked as valid skips, don't know, refused, or not stated/inferred
    - No data in the U.S. national PUF or RUF were imputed
- Assessment Items
    - All missing responses are marked as missing
    - No answers were imputed
- Proficiency scores (in the form of Plausible Values)
    - All were imputed using IRT scaling and the population model

---

Missing Data Codes

Missing Value Codes in SAS and SPSS Datasets		
	SAS missing value code	SPSS missing value code
Valid skip	.V	9-fill ending with 6 (for e.g., 6, 96, 996, …)
Don't know	.D	9-fill ending with 7 (for e.g., 7, 97, 997, …)
Refused	.R	9-fill ending with 8 (for e.g., 8, 98, 998, …)
Not stated or inferred	.N	9-fill ending with 9 (for e.g., 9, 99, 999, …)

More information about missing cases is presented in the module Considerations for Analysis of PIAAC Data

---

Slide 36 of 37

Summary

- The PIAAC sample design
- The weights that must be applied to ensure data are representative of the target population
- The procedures for calculating appropriate standard errors
- How scaling is used in the large-scale international assessments and how plausible values are used when analyzing assessment data
- How to account for missing data ensure accurate analysis

Resources

- Analyzing NCES Complex Survey Data
- Statistical Analysis of NCES Datasets Employing a Complex Sample Design
- Technical Report of the Survey of Adult Skills (PIAAC)
- Data Collected Through the PIAAC
- Program for the International Assessment of Adult Competencies (PIAAC) 2012: U.S. Main Study Technical Report
- Considerations for Analysis of PIAAC Data